**A glimpse of Taiwanese Restaurant**

**-- Din Tai Fung (101 branch)**

**Performance analysis based on food, service, and value for past 7 years**

MING-CHEN, LU

**Section 1 – Motivation and Context**

1. **Motivation**

As an exchange student, I always want to introduce Taiwan to others. Since food is a great start, I decided to choose the famous restaurant in Taiwan – Din Tai Fung. My goal is to understand customers’ preferences of the restaurant and hopefully to get insights for managers of Din Tai Fung in the future.

1. **Problem**

How do customers feel based on food, service, and value after visiting Din Tai Fung for past 7 years? More specifically, is there any dimension or sentiment changes over time?

1. **Approach**

There were four parts in my approach. Python was used in the first part and the rest of them were computed using R.

1. Web Scraping: I scraped entire reviews (3937 reviews) from TripAdvisor dated between October 12, 2011 to September 10, 2018.
2. Data Cleaning: Since using the bag of words approach, I removed numbers, punctuation, whitespace, English stopwords, and convert them to lower cases. In addition, all training texts were randomly sampled to same size (452 words) to compute fairly afterwards.
3. Information Extraction: To perform Naïve Bayes classification, I first calculated the likelihood of each training text. Then, using those likelihood as a source of likelihood for each review. Afterwards, I computed the posterior probability of each review to classify them into different dimensions. As for sentiment analysis, I used the same approach except sentiment analysis was based on each dimension.
4. Analysis: To clearly see the change over time, I will be using word cloud and plots to examine each dimension and sentiment of the reviews.
5. **Domain**

The reason was that eating is an everyday thing and I am most familiar with traditional Taiwanese dishes, which Din Tai Fun is internationally well-known for.

1. **Dimensions**

Three chosen dimensions are food, service, and value. I followed the dimensions on TripAdvisor website and considered them reasonable. While food is the main character when dining, service quality affect customers’ feeling about the restaurant. Overall, value determines the whole performance of the restaurant.

**Section 2 – Training texts**

1. **Training texts** **for each dimension**.
2. Food: The website title is “Taiwanese Food Guide: 57 Things to Eat in Taiwan and Where to Try Them”. Since the author only introduced traditional Taiwanese food and her personal comments in the article, it’s reliable enough. Link: <https://www.willflyforfood.net/2018/04/07/taiwanese-food-guide-57-things-to-eat-in-taiwan-and-where-to-try-them/#xiao-long-bao>
3. Service: A paper that conducts a study of critical service encounters to gain understanding of the particular events that cause customers to distinguish very satisfactory services from very dissatisfactory ones. Since there are many descriptions of customers’ experiences on various services, it’s appropriate to use. Reference: Bitner, M., Booms, B., & Tetreault, M. (1990). The service encounter: Diagnosing favorable and unfavorable. *Journal of Marketing,* *54*(1), 71-71.
4. Value: A chapter that focuses on many variants of values as used to imply degrees of excellence. Reference: Oliver, R. L. (2002). Value as excellence in the consumption experience. In *Consumer value* (pp. 59-78). Routledge.
5. **Training texts** **for each sentiment**.

Since the 300 likelihoods from sentiment lexicons weren’t completely relevant to restaurant, I chose to have my own sentiment.

1. Positive: It came from a website that only contains positive reviews about dining experience. Link: <https://www.kingandprince.com/dining-guest-reviews.aspx>
2. Negative: It came from an article that collected the best of bad restaurant reviews in 2017. Link: <https://www.eater.com/2017/12/27/16795084/bad-restaurant-reviews-2017>
3. **Word count, likelihoods**.

**Content lexicon**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Food | | | service | | | value | | |
| Word | count | likelihood | word | count | likelihood | word | count | likelihood |
| 1 | pork | 35 | 0.0104 | customer | 25 | 0.0075 | quality | 54 | 0.0158 |
| 2 | popular | 32 | 0.0095 | service | 15 | 0.0046 | consumption | 47 | 0.0137 |
| 3 | find | 27 | 0.0081 | diners | 12 | 0.0037 | comparison | 20 | 0.0060 |
| 4 | sweet | 19 | 0.0058 | restaurant | 11 | 0.0034 | terms | 19 | 0.0057 |
| 5 | markets | 18 | 0.0055 | customers | 9 | 0.0026 | price | 15 | 0.0046 |
| 6 | noodles | 18 | 0.0052 | make | 7 | 0.0023 | first | 14 | 0.0043 |
| 7 | sauce | 17 | 0.0049 | excellent | 6 | 0.0020 | values | 14 | 0.0043 |
| 8 | chicken | 16 | 0.0049 | food | 5 | 0.0017 | use | 14 | 0.0043 |
| 9 | dishes | 16 | 0.0049 | problem | 5 | 0.0017 | definition | 13 | 0.0040 |
| 10 | balls | 16 | 0.0049 | experience | 4 | 0.0014 | another | 12 | 0.0037 |
| 11 | just | 13 | 0.0041 | long | 4 | 0.0014 | defined | 12 | 0.0037 |
| 12 | much | 13 | 0.0041 | online | 4 | 0.0014 | individuals | 11 | 0.0034 |
| 13 | get | 12 | 0.0038 | ordering | 4 | 0.0014 | second | 11 | 0.0034 |
| 14 | good | 12 | 0.0038 | serve | 4 | 0.0014 | standard | 11 | 0.0034 |
| 15 | delicious | 12 | 0.0038 | time | 4 | 0.0014 | ideal | 11 | 0.0034 |

Originally, I had 1752,452, and 1600 words for dimension food, service, and value respectively after data cleaning. In order to compute fairly, I randomly sampled all dimension to 452 words. Then, using the likelihood formula described in Zacharski, R. A (2015) Programmer Guide to Data Mining. i.e. let denotes the number of words in lexicon, for each word in the lexicon , compute . By following the formula, I got likelihood of 15 words with the highest count shown in the above table.

**Sentiment Lexicon**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | positive | | | negative | | |
| word | count | likelihood | word | count | likelihood |
| 1 | says | 29 | 0.0119 | restaurant | 18 | 0.0075 |
| 2 | food | 16 | 0.0067 | review | 8 | 0.0036 |
| 3 | great | 15 | 0.0063 | bad | 7 | 0.0032 |
| 4 | guest | 14 | 0.0059 | kill | 7 | 0.0032 |
| 5 | service | 14 | 0.0059 | meal | 6 | 0.0028 |
| 6 | stay | 13 | 0.0055 | new | 6 | 0.0028 |
| 7 | tripadvisor | 13 | 0.0055 | detroit | 6 | 0.0028 |
| 8 | enjoyed | 12 | 0.0051 | killa | 6 | 0.0028 |
| 9 | delicious | 11 | 0.0047 | chicken | 5 | 0.0024 |
| 10 | staff | 11 | 0.0047 | can | 4 | 0.0020 |
| 11 | wonderful | 11 | 0.0047 | dessert | 4 | 0.0020 |
| 12 | anonymous | 10 | 0.0043 | vespertine | 4 | 0.0020 |
| 13 | good | 10 | 0.0043 | like | 4 | 0.0020 |
| 14 | room | 9 | 0.0040 | writes | 4 | 0.0020 |
| 15 | everything | 8 | 0.0036 | terrible | 4 | 0.0020 |

I used the same way as content lexicon to calculate likelihood. The original number of terms were (587,849) for positive and negative sentiment. Then I randomly sampled them to 587 words for each.

**Section 3 – Classification of Reviews**

1. **Website** **with the reviews (or posts or tweets)**

I believe TripAdvisor, the largest travel website in the world, can provide sufficient and relatively reliable user generated content to analyze. Therefore, I chose Din Tai Fung (101 branch) from TripAdvisor. Here is the website link: <https://www.tripadvisor.com/Restaurant_Review-g13808515-d2244808-Reviews-Din_Tai_Fung_101_Branch-Xinyi_District_Taipei.html>

1. **One** **review from the website**
2. Link: <https://www.tripadvisor.com/ShowUserReviews-g13808515-d2244808-r614039910-Din_Tai_Fung_101_Branch-Xinyi_District_Taipei.html>
3. Final posterior probability of the review discussing each dimension:

After cleaning data, I separated review into each word in one row with counts of how many times it appears in a single review. Then, match the words with the likelihood of each dimension. To avoid a very small value, I added the log to the likelihood of term in content. After that, using , where W denotes each term and D denotes each dimension, to compute the posterior probability of each word. Take few word in review for example, the posterior probability of food dimension given {casual} would be . Other posteriors were shown in below table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| prior |  | 0.33 | 0.33 | 0.33 |
| word | count | food | service | value |
| casual | 2 | -5.51688 | 0.00000 | -6.76964 |
| food | 1 | -7.46279 | 0.00040 | 0.00020 |
| just | 1 | 0.00051 | 0.00000 | 0.00071 |

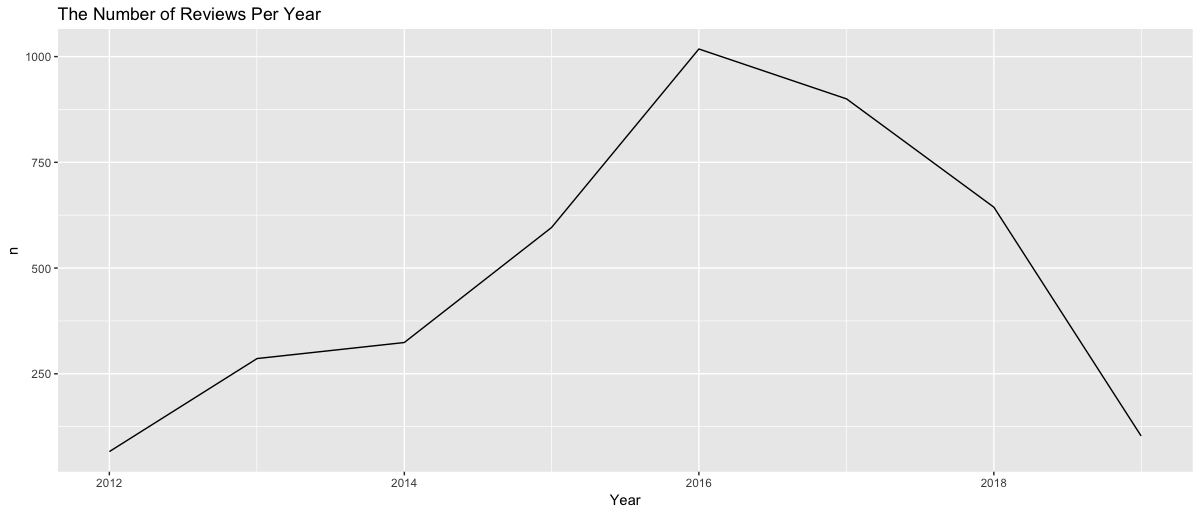
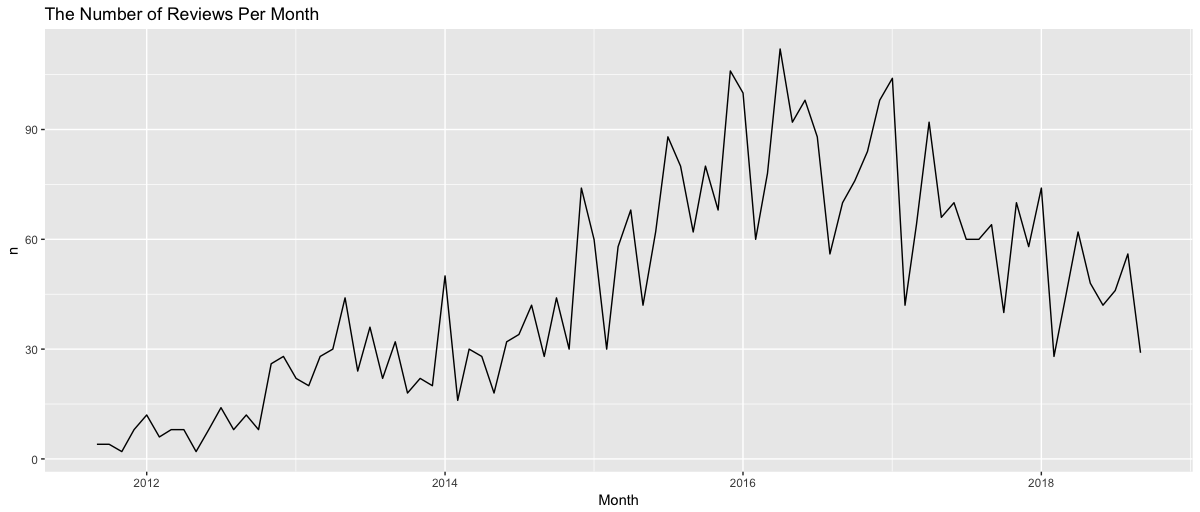
|  |  |  |  |
| --- | --- | --- | --- |
| posterior | food | service | value |
| given {casual} | 0.449018925 | 0 | 0.550981075 |
| given {casual,casual} | 0.39908697 | 0 | 0.60091303 |
| given {casual,casual,food} | 1.000040354 | 0 | -4.03544E-05 |
| given {casual,casual,food,just} | 1.00005618 | 0 | -5.61805E-05 |

The final posterior probability for the review is (0.89901,0,0.10099). With dimension of food having the largest probability, I consider the review mostly about food and classify it into dimension of food.

**Section 4 – Analysis**

1. **Exploratory Data Analysis**

Before classifying each review into different dimension, I attempted to use plots to observe the trend of reviews for past 7 years.

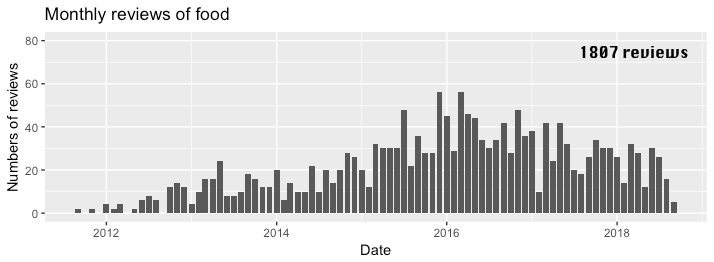
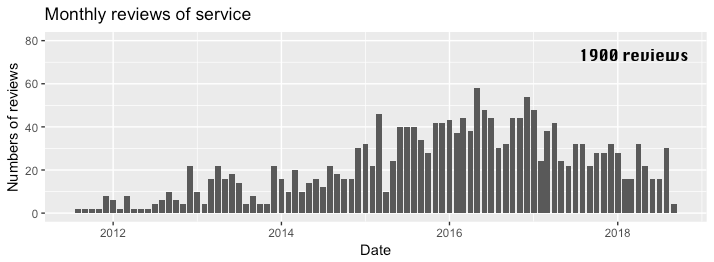
As we can see from the above plots, highest number of reviews were found in 2016 with total of 1018 reviews. The trend might be affected by popularity of the restaurant, numbers of people traveling to Taiwan, world overall economic condition etc. Also, when observing the number of reviews per month, there are fluctuations between months. They may be resulted from peak and off-peak season of the tourism.

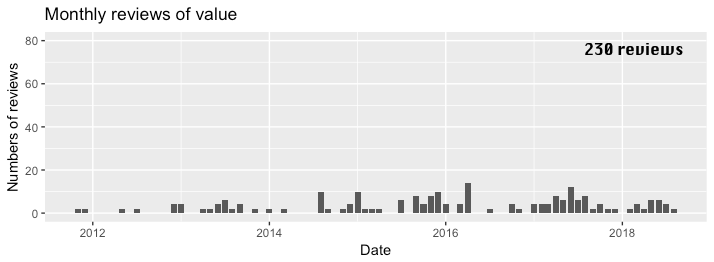
1. **Naïve Based Classification Results**

Dimension1: food Dimension2: service Dimension3: value

**  **

From the word cloud of food dimension, we can see the word “food” appears in high frequency. The word “long” also frequently appears in reviews of the service dimension. Other than that, many words, such as service, good, wait, repeatedly appear in three dimensions. The unpleasing result may be caused by the selection of training text or there is a better way to classify other than naïve based.

**** 

****

The above bar plots show the numbers of reviews per month from 2011 to 2018 after classified into different dimensions. The original purpose was to see whether there was particular mentioning of dimension between months or years so that I got to know changes of customers’ preferences. However, the pattern is similar between dimensions of food and service. And less than 10% of the reviews were classified into dimension of value. From the result, I surmise that classifier is not performing well or food and service are almost equally mentioned in the reviews.

1. **Sentiment Results**

By calculating the sentiment for each dimension, the average positive posterior for food and service were 0.9999 while value was the opposite. The result tended to be unreliable and need further check in the future. Still, in conclusion, the overall sentiment toward Din Tai Fung was positive. There was growing popularity based on the number of reviews per year. Therefore, most customers were satisfied with Din Tai Fung performance over the years.

**Limitations**

1. Due to the results of classification, the quality and preciseness of training text should be examined much more carefully.
2. I only did sentiment analysis at review-level. If I can break down into sentence-level, it is likely that I will get more insights about customers and Din Tai Fung.
3. If I could also successfully scrape down ratings and countries of reviewers, I can construct model to test whether sentiment is related to ratings or countries of reviewers.